DBETH: A Database of Bacterial Exotoxins for Human

Abhijit Chakraborty[1], Sudeshna Ghosh[1], Garisha Chowdhary[2], Ujjwal Maulik[2] and Saikat Chakrabarti[1,*]

[1]Department of Structural Biology and Bioinformatics Division, Indian Institute of Chemical Biology (IICB), Council for Scientific and Industrial Research (CSIR) and [2]Department of Computer Science and Engineering, Jadavpur University, Kolkata, WB 700 032, India

Received August 22, 2011; Revised October 1, 2011; Accepted October 12, 2011

ABSTRACT

Pathogenic bacteria produce protein toxins to survive in the hostile environments defined by the host's defense systems and immune response. Recent progresses in high-throughput genome sequencing and structure determination techniques have contributed to a better understanding of mechanisms of action of the bacterial toxins at the cellular and molecular levels leading to pathogenicity. It is fair to assume that with time more and more unknown toxins will emerge not only by the discovery of newer species but also due to the genetic rearrangement of existing bacterial genomes. Hence, it is crucial to organize a systematic compilation and subsequent analyses of the inherent features of known bacterial toxins. We developed a Database for Bacterial ExoToxins (DBETH, http://www.hpppi.iicb.res.in/btox/), which contains sequence, structure, interaction network and analytical results for 229 toxins categorized within 24 mechanistic and activity types from 26 bacterial genuses. The main objective of this database is to provide a comprehensive knowledgebase for human pathogenic bacterial toxins where various important sequence, structure and physico-chemical property based analyses are provided. Further, we have developed a prediction server attached to this database which aims to identify bacterial toxin like sequences either by establishing homology with known toxin sequences/domains or by classifying bacterial toxin specific features using a support vector based machine learning techniques.

INTRODUCTION

Toxins are one of the most important effector proteins out of many virulence factors utilized by bacterial pathogens to invade and evade the host's immune system. Bacterial toxins are the most powerful poisonous proteins which retain high activity at very high dilutions. Pathogenic bacterial systems have evolved with different types of toxins to counter the defense mechanism of human body (1–4). The cell-associated toxins are referred to as endotoxins whereas the extracellular diffusible toxins are referred to as exotoxins. Exotoxins are usually secreted and act at a site remote from bacterial growth. Exotoxins are usually proteins or polypeptides that act enzymatically or through direct action with host cell receptors to stimulate a variety of host responses (4). Understanding the structural and functional details of bacterial toxins and their relationship with the mode of actions has primary importance in toxoid based therapeutics and bio-defense analyses. With the increasing number of large scale pathogenomics studies the knowledge about the human bacterial pathogens and their effector toxin molecules is rapidly growing. Several databases such as Tox-prot (5), VFDB (6), TvFac (7) and MvirDB (8) exist which aim to collect microbial virulence factor and toxin information. However, these databases mostly concentrate on collection of information from various resources and generally lack subsequent sequence and structure based analyses and predictive protocol for identification of potential bacterial toxins. Hence, it is crucial to maintain a systematic compilation of bacterial toxin sequences facilitating sequence–structure based analyses and classification to understand the molecular details of known bacterial toxins and further utilize the knowledge to identify and/or synthesize potential bacterial toxins.

To address this issue, we developed a Database of Bacterial ExoToxins for Human (DBETH), which not only assembles information of toxins responsible

[*]To whom correspondence should be addressed. Tel: +91 33 2499 5809; Email: saikat273@gmail.com; saikat@csiriicb.in

for causing bacterial pathogenesis in humans but also provides a user interactive platform to perform several sequence and structure based analyses. DBETH has utilized a set of bioinformatics tools and web servers for the analysis of toxin sequences and structures, and has generated useful information regarding bacterial toxins. Subsequently, DBETH has incorporated a server end to identify potential human pathogenic bacterial toxin like protein using sequence similarity based [e.g. BLAST (9)] and support vector machine (SVM) (10) based machine learning approaches. We believe our DBETH database and server will be a useful resource for the biomedical research community aiming to understand and counter the bacterial pathogenesis.

DATABASE IMPLEMENTATION

DBETH is developed on CGI-PERL based web architecture. The architecture implements user friendly web interference for database searching, data visualization and analysis. Server side CGI-PERL action scripts are implied to generate dynamic HTML pages providing interaction within the database resources. Cascading style sheets (CSS) are used for presentation and semantics of the web pages.

DBETH contains sequence, structure, interaction network information and analytical results for 229 toxins categorized within 24 mechanistic and activity types from 26 pathogenic bacterial genuses. A total of 305 experimentally validated three dimensional (3D) structures and 55 *in-silico* modeled 3D structures are available at DBETH database. Table 1 and Supplementary File S1 provide general statistics of the DBETH database. The collection of the toxin dataset was done via a multi stage filtering based approach. In the first stage, existing literature information was collected for the bacterial toxins responsible for pathogenicity in human through PubMed (11) keyword searches (e.g. Bacterial toxins AND Human diseases). In the next step, the names of the individual toxins were collected from the literature (1–4) and searches were made with these toxin names against Genbank database hosted at NCBI (12). A total of 4473 non-hypothetical toxin like sequences were collected from 26 clinically important human pathogens followed by the removal of redundant sequences at the 90% identity threshold using the CD-HIT program (13). The remaining 1143 sequences were mapped to its equivalent UniProt database sequence IDs (14) and were screened on the basis of their Gene Ontology (GO) (15) terms representing pathogenesis and toxicity and other associated virulence terms to select the final 229 toxins enlisted in DBETH. The complete data collection process is depicted in Supplementary File S2.

Table 1. Statistics of DBETH database

Toxins	229
Toxin mechanism types	12
Toxin activity types	12
Bacterial genus	26
Toxin sequences	31 769
Experimental 3D structures	305
In-silico 3D model structures	55
GO localization	219
GO molecular function	173
GO biological process	257
Protein–protein interaction	1186
Domains identified within toxin sequences	338
Motifs identified	260

Following sections briefly describe the browsing and server parts of DBETH.

DBETH database browsing

The main web page of DBETH consists of a list of exotoxins producing pathogenic bacterial genus under the primary category of ‘Pathogens’ listed on the top of the browsing panel, organized on the left side of the web page. Each bacterial genus is dynamically hyperlinked to its respective sources, where the first page is the introduction to the bacterial genus and its respective bacterial species producing the toxins. Details of the toxins can be seen by simply clicking the individual toxin names from the introduction pages. General information along with different physico-chemical properties can be retrieved from the toxins details page. Users are provided with many other resource options, including ‘Domain Organization’ where users can easily visualize the different protein domains present in the toxins. ‘Sequence Alignment’ option enables the users to visualize, edit or download the toxin homologues directly from the respective web page using Jalview applet (16). Phylogenetic tree of the toxins and their respective orthologous sequences is embedded within the Archaeopteryx phylogenetic analysis software (17).

The ‘Structure’ option present in the browsing panel enables users to view the structures of the toxins along with their respective Structural Classification Of Proteins (SCOP) (18) and CATH (19) classifications. Links are provided within the ‘Structure’ page for a given protein data bank (PDB) id (20) to its respective SCOP (18) and CATH (19) classification pages. Not all toxins have the available 3D structures, for such scenario DBETH is providing the ‘Modbase’ (21) link where users can find *in-silico* modeled 3D structure for the toxins. Cases where ‘Modbase’ (21) does not have any modeled structure available, DBETH is providing the 3D modeled structure generated via fold prediction and homology modeling procedure (22). Fold prediction analyses were performed using HH-PRED (23) and PSI-PRED (24) programs while 3D model generation and subsequent loop refinement steps were done with MODELLER 9v8 (25) program.

‘Sequences’ page provides the raw sequences of the toxins for each respective bacterial genus. DBETH is providing an option for a network orthology based approach to identify protein–protein interaction (PPI) information. The ‘STRING’ database (26) was mined to create PPI network for toxins which can be viewed and availed from the ‘Protein Network’ option in the browsing panel bar.

The ‘Sequence and Structure Analysis’ option provides the user a collection of analysis results. It includes analysis of phosphorylation site prediction (27), subcellular localization (28), motif prediction (29), and signal peptide identification (30–31), prediction of peptides binding the MHC molecules (32), structural analysis and primary sequence analysis.

In the DBETH database we are also providing an option where the toxins can be browsed based on their mechanism and activity. A total of 12 such mechanistic types and 12 activity types are defined based on the toxins' biological function.

DBETH server

DBETH is embedded with a server part, where the aim is to identify the potential toxin sequences. The server is divided into two sub parts; the first part includes 'Homology based' toxin identification, which aims to identify toxin specific domains within a given protein sequence using HMMER (33) derived Hidden Markov Model (HMM) profile (34) matching against a toxin domain HMM profile database. The HMM profile dataset is created by running the exotoxins against six different domain database including Pfam-A (35), Pfam-B (35), CDD (36), COG (37), SMART (38) and TIGR (39). Users can also search their sequences against the toxin protein sequences and their homologues using conventional BLAST (9) searching procedure. Users can also search a protein structure against the available DBETH structure database. Structural alignment using Mustang_v3.2.1 (40) enables the user to identify structural similarity within a protein against toxin structures. The second part of DBETH server includes a 'Non-Homology' based approach where a SVM (10) based method is employed to identify potential bacterial toxins. A total of 298 features based on peptide (di-peptide and tri-peptide) frequencies and combinations along with frequencies of amino acids' physicochemical property groups were calculated to characterize the positive (toxins) and negative (non-toxins) samples. LibSVM (41) was used to build the classifier models. A training dataset comprising of 180 bacterial toxins and 1800 non-toxins (1:10 ratio for positive and negative sample) were developed to train the model using *svm-train* program of the LibSVM package. A Radial basis kernel function (RBF) has been used via a 10-fold cross validation of the training set to obtain the optimized gamma (0.5) and C parameter (2.0). Further a feature selection protocol was implemented to remove the possible redundant features from original feature set. LibSVM script *fselect.py* was used to rank the 298 features by assigning them an F-Score value as given by the following equation.

$$F(i) \equiv \frac{\left(\bar{\mathrm{a}}_i^{(+)} - \bar{\mathrm{a}}_i\right)^2 + \left(\bar{\mathrm{a}}_i^{(-)} - \bar{\mathrm{a}}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(a_{k,i}^{(+)} - \bar{\mathrm{a}}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(a_{k,i}^{(-)} - \bar{\mathrm{a}}_i^{(-)}\right)^2}$$

where $\bar{a}_i$, $\bar{a}_{\mathrm{i}}^{(+)}$, $\bar{a}_{\mathrm{i}}{}^{(-)}$ are the average of the *i*th feature of the whole, positive, and negative data sets, respectively; $a_{\mathrm{k,i}}^{(+)}$ is the *i*th feature of the *k*th positive instance, and $a_{\mathrm{k,i}}^{(-)}$ is the *i*th feature of the *k*th negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative. An optimized 114 feature (gamma = 1.0 & C = 4.0) set based on their F-score was selected (please see Supplementary File S3 for a list of optimized features) in the DBETH server to predict the likelihood of a query sequence to be a bacterial toxin.

RESULTS

Figure 1 shows various aspects of DBETH database and server. For example, Figure 1A provides the number of toxins for each bacterial species whereas Figure 1B shows the toxins categorized in multiple mechanistic and activity types. Bacteria containing the largest number of toxins are *Escherichia* (38 toxins), *Clostridium* (29 toxins) and *Staphylococcus* (27 toxins), and the toxins most frequently found are hemolysin (48 toxins), cytotoxin (40 toxins) and those involved in cytoskeletal rearrangements (43 toxins). High numbers of toxins are also found to be involved in proteolysis (31 toxins) and metal binding (28 toxins). With respect to activity, Protease (33 toxins), Phosphatase (28 toxins), ADP-ribosylase (21 toxins) and Lipase (14 toxins) are found to be most abundant. Similarly, Figure 1C–1E show number of toxin 3D structures available for each mechanistic and activity type, SCOP class and bacterial species. Most of the bacterial toxin structures were solved for *Staphylococcus aureus* and *Clostridium botulinum* involved in metal binding proteolytic activity, superantigens and membrane binding/damaging roles. Panels 1F-1H show that majority of the bacterial toxins belong to extracellular localization [Gene Ontology (GO) (14) cellular localization] involved in binding [GO molecular function] and pathogenesis [GO biological process]. We also analyzed the frequency of amino acid types and specific functional groups of amino acids within the bacterial toxin sequences. Interestingly, non-polar (Ala, Gly, Ile, Leu, Val and Pro) and small hydroxyl (Ser and Thr) amino acids are observed to be prevalent in the bacterial toxin sequences (Figures 1I and 1J).

180 toxin sequences and 1800 non-toxin sequences were used as positive and negative dataset, respectively to train the SVM classifier. 5-fold cross-validation simulations were performed using randomly selected 20% dataset as test set. On average (for 5 fold cross-validation) 92.27% accuracy and 0.998 area under curve (AUC) value were obtained when all the features (298) were utilized, whereas 91.16% accuracy and 0.94 AUC value were achieved with an optimized set of 114 features (Supplementary File S4). Higher accuracies (95.54% and 97.21% for 298 and 114 features, respectively) and sensitivities (51% and 71% for 298 and 114 features, respectively) were achieved when an absolutely separate test set consisting of 49 toxins and 490 non-toxins (1:10 ratio) were used to test the classifier's performance (Supplementary File S4). Much better performance of the DBETH SVM based toxin prediction server was observed (Supplementary File S5) when results from 49 toxins and 490 non-toxin sequences were compared against BTXpred server (42), which also predicts bacterial toxins based

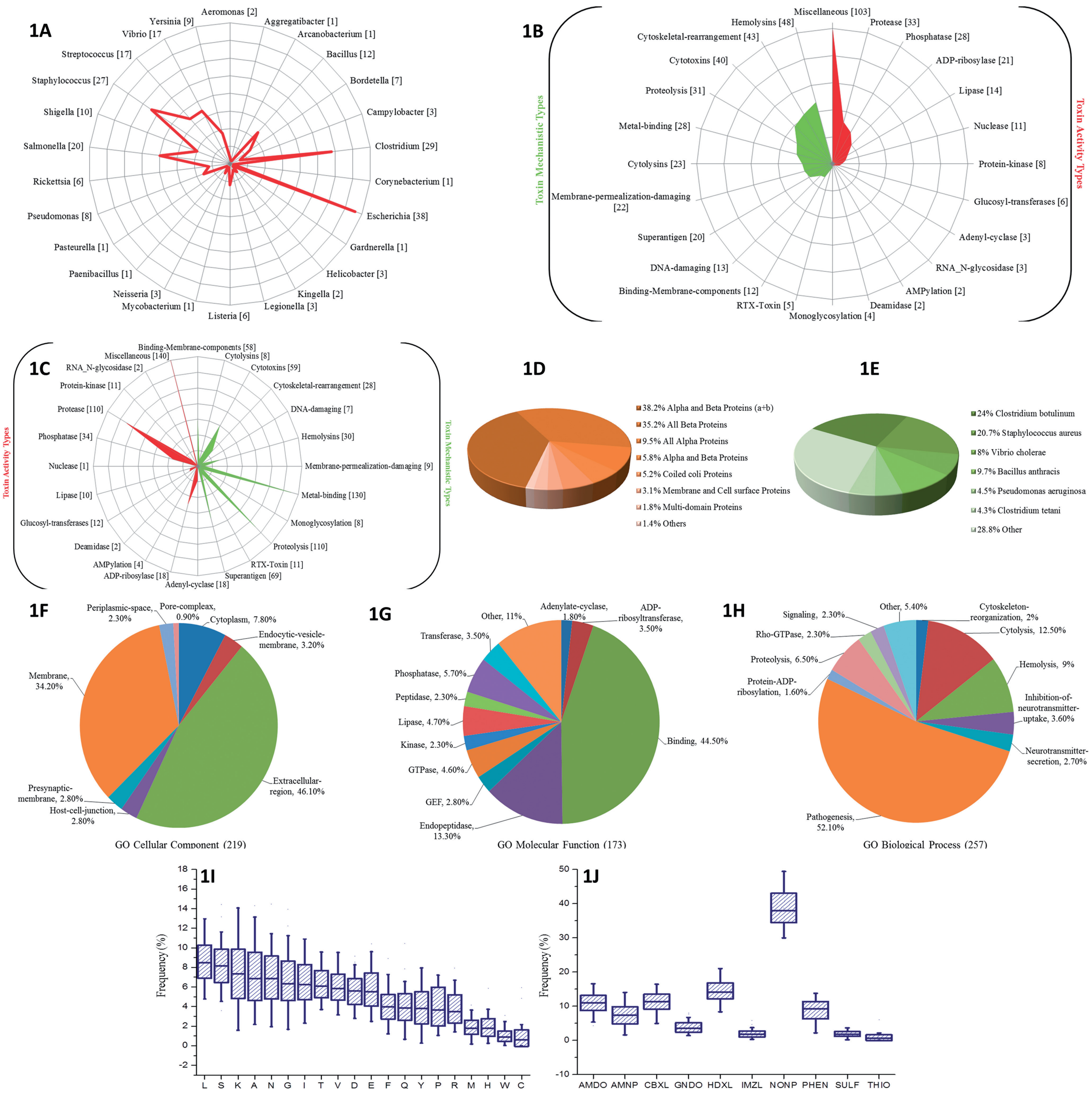

Figure 1. Statistical features of DBETH. (**1A**) Number of toxins in each pathogenic bacterial genus. (**1B**) Number of toxins in each toxin mechanistic and activity types. (**1C**) Number of toxin 3D structures in each toxin mechanistic and activity types. (**1D**) SCOP classification of toxin 3D structures available in DBETH. (**1E**) 3D structure frequencies in pathogenic bacterial genus. (**1F–1H**) Gene Ontology (GO) annotations of cellular localization (1F), molecular functions (1G) and biological process (1H) for bacterial toxins. (**1I–1J**) Frequencies of raw amino acids (1I) and functional groups (1J) within the bacterial toxin sequences. AMDO (Gln and Asn), AMNP (Lys), CBXL (Asp and Glu), GNDO (Arg), HDXL (Ser and Thr), IMZL (His), NONP (Ala, Gly, Ile, Leu, Val and Pro), PHEN (Phe, Trp and Tyr), SULF (Met) and THIO (Cys) stand for Amido, Primary amine, Carboxyl, Guanidino, Hydroxyl, Imidazole, Non-polar, Phenyl, Sulfur and Thiol functional group, respectively.

on primary amino acid sequence. Homology-based detection of toxins using BLAST (9) was also employed to compare the performance of the DBETH SVM based toxin prediction server. A total of 49 toxin sequences were searched against a bacterial specific non-redundant sequence dataset (5 853 363 sequences) after jack-knifing the 49 toxins and their homologous (16 685 sequences) from the dataset. Top BLAST hits (E-value: $\leq$ 1e–05, query-hit alignment length coverage: $\geq$ 50%) were selected as true positive based on their similarity of function with the query toxin protein. Only 27 out of the 49 toxins were matched correctly with the query proteins (Supplementary File S5).

DISCUSSION

Fascinating evolutionary adaptations between human and bacteria result in the development of specific virulence factors and toxins which in turn leads to reorganization of host defense mechanism and immune response. Similarly, this interplay has a significant impact on therapeutic intervention in order to cure bacterial pathogenesis. Over the past decades, our understanding of the mechanisms of action of bacterial toxins has increased enormously due to the advent of large scale pathogenomics studies. These studies have not only identified numerous novel bacterial toxins but also provided insights towards the molecular interactions leading to the discovery of important pathways in cell biology. We have created DBETH to maintain a systematic compilation of bacterial toxin sequences facilitating sequence and structure based analyses and classification to categorize the molecular details of known bacterial toxins. We further utilized the acquired knowledge to identify and/or predict bacterial toxin like sequences using machine learning approaches. We believe that information and resources provided in the DBETH database will be useful in understanding the bacterial toxicology, especially in the design of novel therapeutic strategies (drugs, vaccines and adjuvant) for the management of bacterial toxin-induced diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Files 1–5.

ACKNOWLEDGEMENTS

The authors acknowledge Sayantika Majumdar for her involvement in developing *in-silico* 3D models of bacterial toxins. The authors also acknowledge Preetha Bhattacharjee for her involvement in calculating the features for SVM classification.

FUNDING

Indian Institute of Chemical Biology (IICB), Council for Scientific and Industrial Research (CSIR) to S.C.; Department of Biotechnology for Ramalingaswami Fellowship to S.C. Funding for open access charge: Indian Institute of Chemical Biology, Council for Scientific and Industrial Research, India.

Conflict of interest statement. None declared.

REFERENCES

1. Aktories,K. and Barbieri,J.T. (2005) Bacterial cytotoxins: targeting eukaryotic switches. *Nat. Rev. Microbiol.*, **3**, 397–410.
2. Schiavo,G. and van der Goot,F.G. (2001) The bacterial toxin toolkit. *Nat. Rev. Mol. Cell Biol.*, **2**, 530–537.
3. Ham,H., Sreelatha,A. and Orth,K. (2011) Manipulation of host membranes by bacterial effectors. *Nat. Rev. Microbiol.*, **9**, 635–646.
4. Alouf,J. and Popoff,M. (2005) *The Comprehensive Sourcebook of Bacterial Protein Toxins*, 3rd edn. Academic Press, San Diego.
5. Jungo,F. and Bairoch,A. (2005) Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein knowledgebase. *Toxicon*, **45**, 293–301.
6. Chen,L., Yang,J., Yu,J., Yao,Z., Sun,L., Shen,Y. and Jin,Q. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.*, **33**, D325–D328.
7. Mantri,Y. and Williams,K.P. (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.*, **32**, D55–D58.
8. Zhou,C.E., Smith,J., Lam,M., Zemla,A., Dyer,M.D. and Slezak,T. (2007) MvirDB–a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.*, **35**, D391–D394.
9. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
10. Cortes,C. and Vapnik,V. (1995) 'Support-vector networks'. *Mach. Learn.*, **20**, 273–297.
11. Motschall,E. and Falck-Ytter,Y. (2005) Searching the MEDLINE literature database through PubMed: a short guide. *Onkologie*, **28**, 517–522.
12. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
13. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
14. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
15. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genetl*, **25**, 25–29.
16. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
17. Han,M.V. and Zmasek,C.M. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**, 356.
18. Lo Conte,L., Ailey,B., Hubbard,T.J., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.
19. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.
20. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
21. Pieper,U., Webb,B.M., Barkan,D.T., Schneidman-Duhovny,D., Schlessinger,A., Braberg,H., Yang,Z., Meng,E.C., Pettersen,E.F., Huang,C.C. *et al.* (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **39**, D465–D474.
22. Forster,M.J. (2002) Molecular modelling in structural biology. *Micron*, **33**, 365–384.
23. Soding,J., Biegert,A. and Lupas,A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
24. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
25. Eswar,N., Eramian,D., Webb,B., Shen,M.Y. and Sali,A. (2008) Protein structure modeling with MODELLER. *Methods Mol. Biol.*, **426**, 145–159.
26. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
27. Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.

28. Yu,N.Y., Wagner,J.R., Laird,M.R., Melli,G., Rey,S., Lo,R., Dao,P., Sahinalp,S.C., Ester,M., Foster,L.J. *et al*. (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and pred capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.
29. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
30. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
31. Hiller,K., Grote,A., Scheer,M., Munch,R. and Jahn,D. (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.*, **32**, W375–W379.
32. Reche,P.A. and Reinherz,E.L. (2007) Prediction of peptide-MHC binding using profiles. *Methods Mol. Biol.*, **409**, 185–200.
33. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.
34. Krogh,A., Brown,M., Mian,I.S., Sjölander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
35. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
36. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al*. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
37. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
38. Letunic,I., Doerks,T. and Bork,P. (2008) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
39. Chan,A.P., Pertea,G., Cheung,F., Lee,D., Zheng,L., Whitelaw,C., Pontaroli,A.C., SanMiguel,P., Yuan,Y., Bennetzen,J. *et al*. (2006) The TIGR maize database. *Nucleic Acids Res.*, **34**, D771–D776.
40. Konagurthu,A.S., Whisstock,J.C., Stuckey,P.J. and Lesk,A.M. (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.
41. Chang,C.C. and Lin,C.J. (2011) LIBSVM: a library for support vector machines. *ACM Transact. Int. Sys. Technol.*, **2**, 1–27.
42. Saha,S. and Raghava,G.P. (2007) BTXpred: prediction of bacterial toxins. *In Silico Biol.*, **7**, 405–412.